Examen n°1

preparado por :
Emilien Joly

Este examen esta pensado para durar 2h30. Responder las preguntas con las explicaciones necesarias. Se aconseja dar solamente la información suficiente. Cada ejercicio tiene la ponderación indicada entre paréntesis. La suma de los puntos es 12 que será reportada sin cambio sobre el total de 10. *Ej : Es suficiente de tener la calificación 10/12 para tener la calificación máxima de 10/10.*

1. (4 pts) Estamos suponiendo que una persona registra cada día la información siguiente sobre los días que ha jugado tennis y las condiciones meteorológicas particulares a cada día.

Día	Cielo	Viento	Juega tennis
1	Soleado	Débil	No
2	Soleado	Fuerte	No
3	Nublado	Débil	Si
4	Lluvioso	Débil	Si
5	Nublado	Débil	Si
6	Lluvioso	Fuerte	No
7	Nublado	Fuerte	Si
8	Soleado	Débil	No
9	Soleado	Débil	Si
10	Nublado	Débil	Si
11	Soleado	Fuerte	Si
12	Nublado	Fuerte	Si
13	Nublado	Débil	Si
14	Lluvioso	Fuerte	No
15	Soleado	Fuerte	Si
16	Nublado	Fuerte	No
17	Nublado	Débil	Si
18	Lluvioso	Débil	No
19	Soleado	Débil	No
20	Lluvioso	Fuerte	Si
21	Soleado	Débil	Si
22	Nublado	Débil	No
23	Nublado	Débil	Si
24	Soleado	Fuerte	Si
25	Nublado	Débil	No

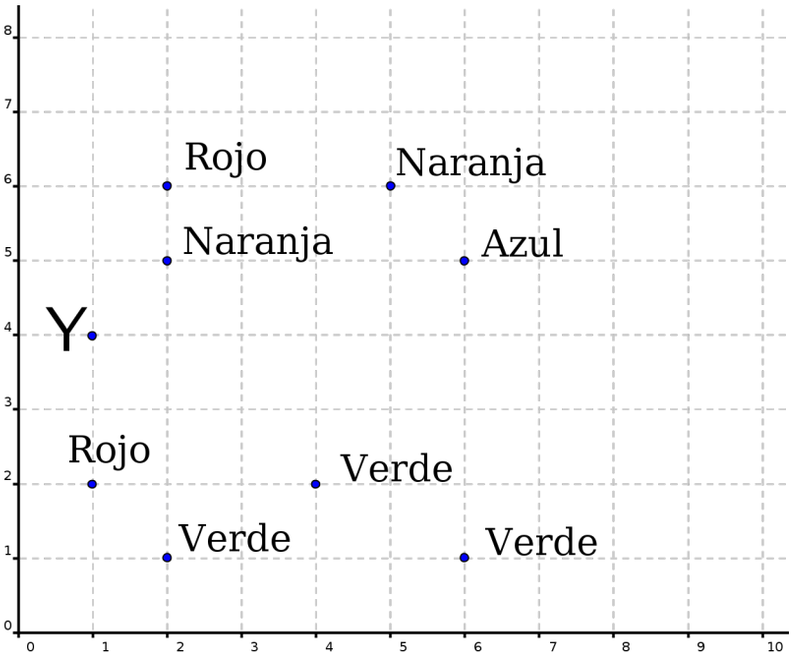
Queremos predecir si la persona jugará tennis los días siguientes
— Día 26 : (Soleado, Viento fuerte)
— Día 27 : (Nublado, Viento débil)
— Día 28 : (Lluvioso, Viento débil)

(a) Proponer un clasificador basado en las ideas de Bayes ingenuo para predecir si la persona jugará tennis esos tres días y dar el resultado de esas predicciones a mano.

(b) Suponiendo que el clima puede cambiar hasta tres veces durante un mismo día, proponer un clasificador que permita predecir la decisión de jugar tennis de la persona en función de la serie de hasta 4 parejas (Estado del cielo, Fuerza del viento) que se pueden ver en el día. Se describirá a detalles como calcular la predicción de un nuevo día.

(c) Proponer un script (en R, Python o Pseudo-code) que permite calcular la predicción del inciso anterior. Verificar que puede (sin calcularlo a mano) predecir por ejemplo los días :
— Día 29 : (Soleado, Viento débil)
— Día 30 : (Lluvioso, Viento débil) , (Nublado, Viento débil), (Soleado, Viento débil)
(Si una función de un paquete está llamada, dar una frase de explicación breve de lo que hace.)

2. (4 pts) Damos el conjunto de datos siguiente con sus etiquetas de colores. La idea es de predecir la clasificación del nuevo punto marcado con la letra Y. En lo que sigue, usamos la notación d para la distancia euclidiana usual.



(a) Dar la predicción del algoritmo k-NN (k-vecinos más cercanos) para todos los valores de k posibles. ¿Que efecto notamos en este ejemplo ?

(b) Proponemos usar una ponderación por pesos en la votación. Imaginemos dar una importancia mayor a los puntos más cercanos de Y. Una propuesta natural es dar el peso $w(x) = e^{-\lambda d^2(x,Y)}$ a un punto x del conjunto de datos. Mostrar que si $\lambda \to \infty$, el clasificador se reduce al 1-NN.

(c) Una otra propuesta de peso es $w(x) = 1/(1 - \lambda d^2(x, Y))$. Mostrar que los dos clasificadores ponderados son equivalentes cuando $\lambda \to 0$.

(d) Proponer un script (en R, Python o Pseudo-code) que permite calcular el clasificador del inciso (c). (Supongamos dado una función `distancia`(a,b) que calcula la distancia euclidiana entre dos puntos a y b)

3. (4 pts) **Un poco de geometría** (Consejo : hacer dibujos !)
Consideramos el conjunto de datos donde cada punto es uno de los vertices del hipercubo de dimensión d dado por $H = \{-1, 1\}^d$. Un punto es negro si tiene el mismo numero de -1 que de 1 en su escritura vectorial. Los otros puntos del hipercubo son blancos.

(a) Calcular $\sum_i x_i$ para los puntos negros. ¿Que valores toma esta suma para los punto blancos ?

(b) Mostrar que en este caso $L^* = 0$ y exhibir el clasificador g^*. ¿g^* es un clasificador lineal ?

(c) Mostrar que un punto y su simétrico obtenido por simetría central por el origen son de mismo color. Deducir que $L \neq 0$ donde L es el riesgo mínimo dentro de los clasificadores lineales.

(d) Consideramos el conjunto $\mathcal{G}_2$ de los clasificadores obtenidos con la discriminación gracias a dos hiperplanos. Mostrar que sobre nuestro conjunto de datos,

$$\min_{g \in \mathcal{G}_2} L(g) = 0.$$